EDGE-Skills

Data Scientist Training

Trainer Script

Trainer Script: Companion material for the presenter

31 slides · vs10.5 · July 2026

This script accompanies the trainer slide by slide with background knowledge, context, and talking points that go beyond the keywords shown on the slides.

Co-funded by the European Union

Slide 1

Title Slide: EDGE-Skills Data Scientist Training

Welcome to the EDGE-Skills Data Scientist Training. Over the next 90 minutes, you will learn how to work with data inside a data space: how to discover data, access it, transform it, and re-offer your results as new assets in the ecosystem.

We use the Prometheus-X open source stack and the EDGE-Skills data space as a concrete example. The focus is on the complete data path from ingestion to value creation, always with an eye on semantic standards, governance, and data quality.

Slide 2

Overview: What This Training Is About

Four guiding principles for this training. First, the data journey approach: we trace how data flows from its source through all transformation stages to value creation. Second, the data scientist focus: how you concretely access data, transform it, and feed results back into the ecosystem.

Third, semantic-driven thinking: standards, ontologies, and metadata are the key to interoperability. Without semantic alignment, data from different organisations simply cannot be combined. Fourth, a practical reference case: we follow a real scenario from raw data to finished insights.

The Skill Gap Analytics use case will accompany us throughout.

Slide 3

Foundation: What Is a Data Space?

The definition is based on the DSSC Blueprint (Data Spaces Support Centre) and the emerging ISO/IEC 20151 standard. Seven elements define a data space: environment rather than platform, trust through verifiable credentials, peer-to-peer communication via connectors, a governance authority that maintains rules, the DSP and DCP protocols plus the PDI (Personal Data Intermediary), machine-readable ODRL policies, and supporting services.

On the right side: what a data space is NOT: not a data lake, not a central platform, not an API gateway, not blockchain-dependent, not a centralised identity provider. For data scientists this means: you build components in a protocol-driven ecosystem, not a monolithic application.

Your pipelines need to work with standards and policies.

Slide 4

Why This Matters for Data Scientists

Four paradigm shifts at a glance. From siloed databases to a federated data ecosystem: you no longer work with a single database but with data from many organisations. From manual data requests to automated, policy-based data access: no more email ping-pong, the connector handles access.

From copying and transforming locally to sovereign data sharing with provenance: data is not copied and processed locally but flows in a controlled manner with full traceability. From static CSV exports to semantic standards and catalogue discovery: you discover data through the catalogue with semantic search rather than via spreadsheet handovers.

Slide 5

The Prometheus-X Architecture: Five Building Blocks

The Prometheus-X open source stack consists of five building blocks and a consent-driven approach centred on the PDI. The five building blocks: PDC Connector for peer-to-peer data exchange, Data Planes where data actually lives, Identity Hub for credential management, Catalogue for federated discovery, and Contract Service for automated agreements.

Prometheus-X implements the DSP (Dataspace Protocol) as an interoperability protocol for discovery, negotiation, and transfer. The key differentiator is the PDI (Personal Data Intermediary), a consent-driven personal data sharing service where individuals manage their consents from a single place. DCP (Decentralised Claims Protocol) handles credential-based identity verification. Service chains enable multi-participant data flows orchestrated across organisations.

The components are connector (PDC), data planes, identity hub, catalogue, and contract service.

Slide 6

The Data Journey in a Data Space

The data journey in five steps, and the central point: collection does not equal value creation. Step 1: data is collected from diverse sources with provenance. Step 2: semantic standards make data interoperable across organisations.

Step 3: transformation and harmonisation prepare data for analytics. Step 4: quality assurance ensures trustworthiness and veracity. Step 5: results are shared back as new data offerings in the ecosystem. Every step in the data journey adds value and builds trust.

This is the core idea that accompanies you through the entire training.

Slide 7

Data Flow in Practice

This slide shows the concrete data flow across three layers. The data layer: assets from multiple sources with Dataspace Connectors (PDC), metadata and usage conditions flowing into the hub. The transformation layer: data is transformed, harmonised, schema-mapped, and run through quality assurance pipelines.

The analytics and application layer: this is where actual value creation happens: supply chain analytics, feasibility analyses, recommendations, and trade-off analyses. The hub connection enables federated discovery and sovereign data exchange across organisational boundaries.

The data stays at the source; only metadata and results flow.

Slide 8

Data Flow Diagram

This slide shows a graphical representation of the data flow. Take a moment to study the diagram. It visualises how data flows from sources through the data space hub, through transformation and quality checks, to the analytics engine.

The arrows indicate the direction of data flows. Note that the data itself remains at the sources; what flows through the hub is metadata, policies, and controlled access authorisations. The actual data transfer happens peer-to-peer through the data planes.

Slide 9

Reference Case: Skill Gap Analytics across Organisations

Our use case: Schülerkarriere as data provider shares anonymised student skill profiles. The Skill Analytics Service as analytics layer maps skills to ESCO and identifies competency gaps. imc as data consumer uses the results for designing training programmes.

The data flow model: data flows through semantic transformation and quality checks. Each step adds metadata and provenance. Results are re-offered as new data assets with their own usage conditions. This is the complete cycle: from raw data to actionable insights and back into the ecosystem.

Slide 10

Step 1: Data Ingestion & Connectors

How data enters the data space ecosystem. Data sources are connected via connectors: databases, APIs, file systems, IoT streams. Dataspace Connectors (PDC) provide standardised access. Each dataset becomes an asset with metadata, a DataAddress, and usage conditions that define who can access it.

Three ingestion patterns: pull for queries on demand, push for batch deliveries, stream for continuous data flows. The central principle: data stays at the source.

Only metadata and usage conditions are shared via the catalogue.

Slide 11

Step 2: Semantic Standards & Ontologies

Without shared semantics, data from different organisations cannot be compared, combined, or analysed. Semantic alignment is the foundation of every successful data space. Three core standards: xAPI for learning experience traces and activity statements.

JSON-LD for machine-readable semantic context as linked data. RDF for knowledge graphs and ontology relationships. Classification frameworks: ESCO as the European standard for skills, competences, and qualifications. ROME as the French job classification framework.

Plus custom ontologies for domain-specific extensions. Your task as data scientists: prepare data so that it can be linked across organisations via semantic relationships.

Slide 12

Step 3: Data Transformation & Preparation

The transformation pipeline in four steps. Map: schema alignment and ontology mapping, your data is aligned to shared standards. Transform: format conversion and normalisation. Enrich: semantic enrichment and metadata addition.

Validate: quality checks and verification. Two core principles: first, quality checks at every pipeline stage, with metadata tracking provenance and lineage. Second, build reusable transformation services that other participants can chain.

This is aligned with the Prometheus-X transformation services. Your pipelines are not one-off scripts; they become services in the ecosystem.

Slide 13

Step 4: Metadata & Catalogue Management

For data to be discoverable, understandable, and reusable, you need three types of metadata. Technical metadata: format, schema, size, encoding, update frequency, API endpoints, data lineage and provenance. Semantic metadata: ontology references such as ESCO and ROME, domain classification and tags, semantic relationships to other assets.

Contractual metadata: usage conditions as ODRL policies, licence type and access restrictions, consent requirements and scope. The chain is: rich metadata enables discovery, discovery builds trust, trust enables reuse. Everything is registered in the Prometheus-X catalogue and enables federated discovery across data spaces.

Slide 14

Step 5: Data Quality & Veracity

Four quality dimensions: accuracy: correctness of values. Completeness: no missing fields. Timeliness: data is current. Consistency: no contradictions. Provenance and lineage: track where data comes from, how it was transformed, and who processed it.

The Data Value Chain Tracker provides end-to-end visibility across the entire data journey. In multi-party environments with multiple organisations: source trust verifies provider reliability. Validation rules automate quality checks.

Monitoring enables continuous quality tracking. Remediation creates feedback loops back to providers. Prometheus-X Data Veracity Assurance validates data integrity at each step.

Slide 15

Step 6: Governance & Policies for Data

Governance shapes what data you can access and share. Three governance layers: access policies define who can discover and access your data assets (open data for all members, restricted only for verified research institutions).

Usage policies determine what can be done with data after access (purpose-limited, for example analytics only, no redistribution). Access policies can also be time-limited (for example access expires after 90 days). Consent management gives human-centric control over personal data, required for any individual-level data exchange under GDPR.

ODRL as Open Digital Rights Language defines machine-readable policies that connectors enforce automatically. As a data scientist, you must understand the policies that govern your data before building analytics pipelines.

Slide 16

Step 7: Publish Data & Re-Offer Results

In three steps, your data and analytics results become available. First: register data asset: define metadata such as type, domain, schema, quality level, and provenance. Point to where the data actually lives. Second: set usage conditions: define who can access, for what purpose, and under which constraints.

ODRL policies are enforced automatically. Third: re-offer results: your analytics results become new assets. Publish insights back to the catalogue for others to discover and reuse. The central principle: the actual data stays on the provider's data plane.

The control plane holds only the governance information.

Slide 17

Step 8: Consume & Integrate Data

The data scientist workflow in a data space in five steps: discover relevant datasets, check quality and provenance, negotiate access automatically, integrate into your pipeline, re-offer enriched results. Technically: discover via the provider's DSP endpoint, filter by semantic tags, quality level, or data type.

Negotiate access: automated policy check, credentials verified, contract agreed. This takes seconds, fully automated. Integrate and analyse: data is received via secure transfer and fed into your analytics pipelines. The data space gives you standardised access to data across organisational boundaries with built-in governance.

Slide 18

VisionsTrust: Your Data Access & Consent Platform

VisionsTrust is the reference implementation of the Prometheus-X open source stack. Three services: the catalogue for browsing available datasets, discovery across organisations, viewing metadata and samples. The contract service for automated data agreements with policy-based access control, no manual approvals needed.

The consent service for personal data governance with user-level consent flows, enabling GDPR-compliant data exchange. These are the three services through which you as data scientists interact with the data space.

Slide 19

How Data Scientists Access Data

The complete flow from discovery to data in your notebook. Five steps: discover: browse the VisionsTrust catalogue for datasets and service offers across organisations. Negotiate: contracts are negotiated automatically, policies evaluate your organisation's credentials.

Consent: for personal data, user consent is collected via the PDI consent service before exchange. Exchange: data flows peer-to-peer through the PDC, you receive it via endpoint and access token. Analyse: ingest into your pipeline: Pandas, Spark, notebooks.

Sovereign data, your familiar tools. The point: you do not need to re-learn. Your analytics tools stay the same; the access to data becomes standardised and governance-compliant.

Slide 20

Resources & the Catalogue: What You See When Browsing

When browsing data in the VisionsTrust catalogue, you see four core fields. Name and type: display name and classification as data or service resource. Description and location: what the data contains and where it is geographically stored, relevant for compliance checks.

Sample extract: a JSON sample showing data structure so you can assess the schema before requesting access. Personal data flag: marks whether consent requirements apply. If flagged, user consent is required before any exchange.

These four fields help you quickly evaluate whether a dataset is relevant and accessible for your analysis.

Slide 21

Four Data Exchange Protocols

Which protocol applies depends on the data type and use case. B2B non-personal: direct organisation-to-organisation data sharing, contract-based authorisation only, no consent needed. Consent-driven: individual consent required, endpoint must include userId, resource flagged as personal data.

API consumption: data provider consumes services, resource marked as API payload, response sent back to provider. Service chains: multi-participant workflows with data plus service plus infrastructure, requires PDC v1.9.0+.

For the skill gap use case, we use consent-driven because student profiles contain personal data.

Slide 22

Consent Flow for Personal Skills Data

When your analysis requires individual-level data, the consent flow kicks in with five steps. The provider flags the resource as personal data in the VisionsTrust metadata configuration. The consent service automatically generates a privacy notice for the data subject (in our case the learner).

The individual grants consent via the PDI consent interface before any data flows. The PDC receives consent confirmation and authorises the secure peer-to-peer data transfer. Personal data flows to your endpoint with the userId parameter for per-user data ingestion.

This flow is not optional; it is a GDPR requirement for personal data.

Slide 23

Testing Your Data Flows

VisionsTrust offers a Tech Space for validating data exchanges before production. Exchange tests: validate all four protocol types: project exchange, service chain, API consumption, consent-driven. Ping connector: verify PDC connectivity and check that your data connector responds correctly to requests.

Log viewer and export: monitor exchange logs in real time and download CSV exports for detailed troubleshooting. Common errors: missing tech config: set REST source and endpoint URL. Missing PII declaration: enable personal data flag and add userId parameter.

Test thoroughly before going to production.

Slide 24

Operational Interfaces

An overview of comprehensive tools for managing data spaces in production. Eight areas: data catalogue for asset registration and classification. Data spaces for browsing federated spaces and participants. Connections for agreement negotiation and exchange management.

Transfers for monitoring data movement and status. Compliance for DPP, CSRD, AI Act, LkSG, EUDR, and DGA. Governance for classification, policy enforcement, and quality scores. Policy engine for multi-licence management and conflicts.

Graph view for network visualisation of the ecosystem.

Slide 25

Privacy-Preserving Data Processing

Personal data in data spaces requires special handling. Privacy-preserving techniques enable analytics while protecting individuals. Three main techniques: anonymisation removes identifying information permanently. Pseudonymisation replaces identifiers with reversible tokens.

Synthetic data creates statistical equivalents. Three advanced approaches: federated processing computes directly at the data source. Differential privacy adds calibrated noise. Secure computation enables multi-party computation.

The core principle: balance utility and privacy. Choose the right technique for each use case. For the skill gap use case, Schülerkarriere uses anonymisation of profiles before sharing.

Slide 26

Validate the Full Data Flow

Six validation steps for the end-to-end check. Ingestion: data sources connected and accessible. Semantics: standards applied, data interoperable. Quality: accuracy, completeness, provenance verified. Governance: contract agreement properly configured.

Analytics: pipeline produces valid insights. Re-offering: results published as new assets. The data flow checklist: can data be discovered via catalogue with semantic search? Are semantic standards correctly applied (xAPI, JSON-LD)? Does the contract agreement get created successfully (FINALIZED)? Are usage conditions and consent properly enforced? Are audit records generated for governance and compliance tracking?

Slide 27

Where Data Integration Often Fails

Six typical failures, and these are not edge cases but the most common reasons data integration in data spaces fails. Ingesting data without applying semantic standards. Skipping metadata enrichment and catalogue registration.

Ignoring data quality checks and provenance tracking. Treating transformation as a one-off script instead of a reusable pipeline. Forgetting consent and usage conditions for personal data. Building analytics in isolation without re-offering results.

Each of these points is a project killer. Go through the list and honestly check whether you have covered all points in your own context.

Slide 28

The Data Scientist Toolkit

Three tool categories you need to know. Standards and semantics: xAPI, JSON-LD, RDF as data formats. ESCO and ROME as classification frameworks. ODRL for machine-readable policies. Linked data principles as design philosophy.

Platforms and tools: VisionsTrust Platform for catalogue browsing, Prometheus-X catalogue for discovery, Dataspace Connectors (PDC) for data access, data plane SDKs for integration. Quality and governance: data veracity assurance for integrity checks, value chain tracker for provenance, consent management via VisionsTrust, PDC Governance and Quality API.

The core principle: invest in semantic alignment and data quality. The infrastructure handles discovery, trust, and transfer automatically.

Slide 29

Exercise: Apply This to Your Data Context

Now it is your turn. Five questions for your own data context: what data do you have: inventory your assets, formats, schemas, and quality levels. What standards apply: map your data to semantic standards like xAPI, ESCO, or JSON-LD.

What transformations are needed: design reusable pipelines for mapping, enrichment, and quality. What governance applies: define usage conditions, consent requirements, and privacy requirements. What results can you re-offer: plan how your analytics outputs become new data offerings in the ecosystem.

Take five minutes and sketch out answers for your own work context.

Slide 30

Key Takeaways: What to Remember as a Data Scientist

Five key messages. First: semantics are the foundation: without shared standards, data from different organisations cannot be combined. Second: quality creates trust: data quality, provenance, and lineage are what make data reusable.

Third: governance enables sharing: policies and consent make it safe to share sensitive data. Fourth: re-offering completes the cycle: your analytics results become valuable assets for others. Fifth: think in data flows, not databases: the data journey model: data, transformation, analytics, application.

Thank you for participating, questions?

Slide 31

Resources & Next Steps

To close, the key resources for your continued work. The EDGE-Skills curriculum, specifically Module 4 on data and semantics (that is your data-deep module). The Prometheus-X documentation for building blocks, standards, and architecture.

The VisionsTrust Platform for catalogue, governance, compliance, and data exchange. The prometheus-x.org site for learning paths, concepts, and hands-on guides. And the Prometheus-X reference implementation at prometheus-x.org for the 4-layer data flow: data, transformation, analytics, application.

These resources form your reference for practical implementation.